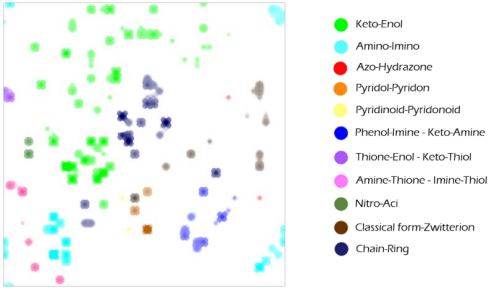
M. Glavatskikh [1,2]
T. Madzhidov [2]
R. Nugmanov [2]
T. Gimadiev [1,2]
D. Horvath [1]
G. Marcou [1]
A. Varnek [1]

QSPR MODELING OF TAUTOMERIC EQUILIBRIA USING LOCAL DESCRIPTORS

[1] Laboratoire de Chémoinformatique, UMR 7140 CNRS, Université de Strasbourg, 1, rue Blaise Pascal, 67000 Strasbourg, France ;
[2] Laboratory of Chemoinformatics and Molecular Modeling, Butlerov Institut of Chemistry, Kazan Federal University, Kremlevskaya 18, Kazan, Russia;
mvglavatskikh@gmail.com

Existing tools predict the ratio of tautomers in aqueous solution at room temperature using the pKa values of related tautomers. This may significantly affect the accuracy, especially, if the errors of the pKa predictions are compareble with the difference of tautomers' pKa values. The problem can be solved if modeling is performed directly for equilbrium constants of tautomeric equilibrium (logK) in solution.

Here, the models were built on a data set of 704 reactions, devided into 11 tautomeric classes, for which logK values were measured in different solvents and at different temperatures[1]. The models were prepared using Support Vector Machine[2] (SVM) and Generative Topographic Mapping[3] (GTM) on ISIDA Fragments[4] and EED descriptors[5]. Both of these descriptor types were enhanced by adding special descriptors for solvent and temperature. In SVM calculations, the Consensus Model was composed of five best individual models (RMSE=0.66-0.69 R^2=0.79-0.83) based on atom-centered fragments of ISIDA descriptors. This model reasonably performs on two external test sets that include the reactions under new reaction conditions (*test 1*) or new structures (*test 2*) and contain, respectivly, 23 and 21 tautomeric equilibria (RMSE=0.59 and 1.25, R^2=0.75 and 0.77). Large RMSE value for *test 2* is explained by the fact that more than half of the compounds were out of the models applicability domain. The consensus model is publicly available on our web-server: https://cimm.kpfu.ru/development/predictor



GTM represents the equilibria as data points projected on 2-dimensional map, on which all 11 classes are well separated (see Figure 1). The GTM-based regression model performs similarly to the SVM model.

Figure1. Classification GTM map based on ISIDA descriptors.

1. Palm V. A.VINITI: Moscow, 1978.
2. Chang C-C. et al. *ACM Trans. Intell. Syst. Technol.*, 2011, **2 (3):** 1-27.
3. Varnek A. et al. *J. Comput.-Aided Mol. Des.*, 2005, **19 (9-10):** 693-703.
4. Braban M. et al. *J. Chem. Inf. Comput. Sci.*, 1999, **39 (6)**: 1119-1127.
5. Kireeva N. et al. *Ind. Eng. Chem. Res.*, 2012, **51 (44)**: 14337-14343.

The research was supported by Russian Scientific Foundation, grant 14-43-00024